This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2851670, IEEE Transactions on Image Processing

Deep Visual Discomfort Predictor for Stereoscopic 3D Images

Heeseok Oh, Sewoong Ahn, Sanghoon Lee, *Senior Member, IEEE* and Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*—Most prior approaches to the problem of stereoscopic 3D (S3D) visual discomfort prediction (VDP) have focused on the extraction of perceptually meaningful handcrafted features based on models of visual perception and of natural depth statistics. Towards advancing performance on this problem, we have developed a deep learning based VDP model named Deep Visual Discomfort Predictor (DeepVDP). DeepVDP uses a convolutional neural network (CNN) to learn features that are highly predictive of experienced visual discomfort. Since a large amount of reference data is needed to train a CNN, we develop a systematic way of dividing S3D image into local regions defined as patches, and model a patch-based CNN using two sequential training steps. Since it is very difficult to obtain human opinions on each patch, instead a proxy ground-truth label that is generated by an existing S3D visual discomfort prediction algorithm called 3D-VDP is assigned to each patch. These proxy ground-truth labels are used to conduct the first stage of training the CNN. In the second stage, the automatically learned local abstractions are aggregated into global features via a feature aggregation layer. The learned features are iteratively updated via supervised learning on subjective 3D discomfort scores, which serve as ground-truth labels on each S3D image. The patch-based CNN model that has been pretrained on proxy ground-truth labels is subsequently retrained on true global subjective scores. The global S3D visual discomfort scores predicted by the trained DeepVDP model achieve state-of-the-art performance as compared to previous VDP algorithms.**

***Index Terms*—Stereoscopic 3D, visual discomfort prediction, convolutional neural network, proxy ground-truth label**

I. Introduction

Stereoscopic 3D (S3D) provides an enhanced sense of reality by enabling the perception of depths on two-dimensional displays. This is accomplished by enabling viewers to experience 3D visualizations by introducing binocular disparities between images presented to the left and right eyes. However, the S3D viewing experience is not always an optimal one, and can be physically uncomfortable. One important reason for this arises when problematic 3D input stimuli are viewed that produce abnormal cross-coupled interactions between the oculomotor and crystalline lens control systems, causing sensations of visual discomfort in the viewer [1][6].

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2016R1A2B2014525).

H. Oh, S. Ahn and S. Lee (corresponding author) are with the Department of Electrical and Electronics Engineering, The Yonsei University, Seoul 120-749, Korea, (e-mail: angdre5@yonsei.ac.kr; anse3832@yonsei.ac.kr; slee@yonsei.ac.kr).

Alan C. Bovik is affiliated with Laboratory for Image and Video Engineering (LIVE), the Department of Electrical & Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084 USA.

A. Related Works and Limitations

A variety of factors have been identified that produce feelings of visual discomfort while viewing S3D content, including crosstalk, object compactness, keystone effects, window violations, and optical distortions, among others [2]-[5]. However, it is widely believed that the most significant causes of visual discomfort are neuronal and oculomotor conflicts arising from accommodation and vergence mismatches (AVM), which are often associated with excessive disparities [6]-[9].

Several visual discomfort prediction (VDP) models that are based on estimating the effects of AVM from S3D images have been proposed. Most of these have focused on the extraction and pooling of perceptually relevant features, that along with recorded subjective discomfort opinion scores, are used to train a regressor to predict discomfort levels. Early VDP models have focused on statistical features descriptive of the distribution of disparities (e.g., average and variance) [10]-[15]. More recently, advanced VDP models based on models of human visual perception have been proposed. Jung *et al*. [16] developed a saliency-based VDP model, whereby saliency-weighted disparity and disparity gradient features were extracted from computed disparity maps. The authors of [17] developed a VDP model that uses a model of retinal resolving power and basic principles of physiological optics. By formulating expressions describing both 2D and 3D visual bandwidths, they defined four types of discomfort features that are predictive of two types of AVM anomalies: absences of defocus blur and absences of differential blur. The model in [7] was designed to extract features sensitive to interactions that occur between the accommodation and vergence mechanisms, the level of out-of-focus blur, and the degree of diplopia. An algorithm was then learned which predicts their collective effects on experienced visual discomfort.

One problem with these approaches is that the VDP models that are created [10]-[17] rely heavily on "handcrafted" feature representations which, even if they capture key perceptual or neurophysiological contributions to 3D discomfort, may yet miss other import latent factors. Generally, it is a challenging problem to adequately model the optical, psychophysical, and physiological elements that are implicated in causing discomfort. Even if the principal factors that cause visual discomfort were revealed, it is difficult to determine how to correctly weight and combine features representative of these factors during the learning process, since the feature extraction, aggregation, and regression sub-processes are functionally independent.

B. Motivation

Towards advancing progress on this complex problem, we have developed a novel deep visual discomfort predictor (DeepVDP) that utilizes a convolutional neural network (CNN) [18][19]. DeepVDP accurately conducts S3D VDP by learning features and abstractions from labelled S3D images and disparity maps computed on them. While deep learning has recently been applied to the image quality problem [20], we believe that this is the first attempt to apply deep learning methods to the VDP problem. We have designed the model structure based on established physiological models of binocular perception, which are used in the construction of each stage of the system.

However, generating the sufficiently massive amounts of subjective data that is required to train a CNN without overfitting is very difficult in this context. Human subjective 3D discomfort labels must be obtained in a controlled laboratory setting and cannot be crowdsourced. Existing VDP datasets contain insufficient amounts of subjective data to be able to meaningfully train even a moderately deep network. For example, the IEEE-SA database [21], which is the largest S3D visual discomfort database, contains only 800 S3D image pairs, which is much smaller than datasets used in typical deep learning applications [24], [25]. Most of the data augmentation techniques that are commonly used to surmount insufficient data volumes, such as image rotations, cropping and vertical flips cannot be used to overcome this problem, since the experienced degrees of visual discomfort are likely to be modified by most geometric transformations of an S3D image.

One appealing idea to capture more data would be to exploit video discomfort datasets. This is highly problematic, since the limited datasets available only supply whole-video labels but no frame scores which are not acceptable since discomfort factors change over time. More importantly, even if an S3D video discomfort database with frame-level scores were available (there is not), it could not be used since powerful motion-related discomfort factors (very fast or chaotic 3D motions) would seriously bias the subjective discomfort labels. Among the various existing augmentation schemes, we have only found swapping the left and right images, followed by horizontally flipping them (to preserve physically correct depth impressions) to be useful.

C. Proposed Approach

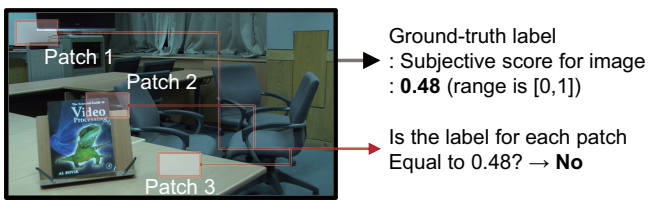


Fig. 1. The subjective MOS is 0.48 for the S3D image "ISS7_50" in the IEEE-SA dataset. However, the ground-truth labels for each patch should not necessarily match the globally assigned subjective score.

We have taken a different approach to overcome the lack of labelled S3D image data. Specifically, we have developed a patch-based CNN model whereby each S3D image used for training or testing is first partitioned into patches. Patch-based learning of a VDP model is problematic, since ground-truth labels are not available for each patch. The whole-image subjective labels cannot be used, since global experiences of visual discomfort are unlikely to coincide with local S3D discomfort contributions. We depict this problem in Fig. 1, where the normalized MOS of the example S3D image ("ISS7_50" in the IEEE-SA dataset) is 0.48. However, when the image is divided into patches, it is difficult to justify assigning the same MOS value to every patch (e.g., patches 1–3 in Fig. 1). The most reliable method to obtain the local level of visual discomfort, of course, would be by collecting subjective scores on each patch; this, however, is not a reasonable possibility [26].

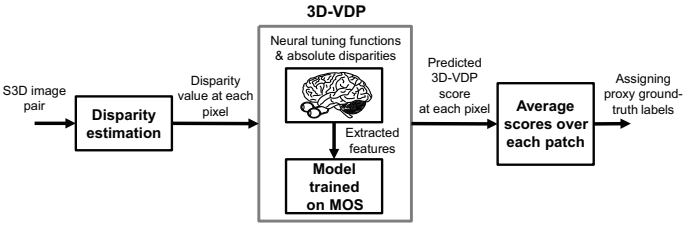


Fig. 2. **Proxy ground-truth labels are generated using the 3D-VDP algorithm, and then assigned to the corresponding image patches.**

To cope with this, we introduce a new concept of learning on *proxy ground-truth patch labels* which are used in lieu of patch subjective scores. The proxy ground-truth labels are obtained on each patch in the form of the responses of an S3D VDP algorithm (called 3D-VDP) [8].

Fig. 2 depicts the system used to assign proxy ground-truth labels generated by 3D-VDP to patches of a S3D image. Generating proxy ground-truth labels for each patch consists of three steps: training the 3D-VDP model over entire images, predicting a 3D-VDP score at each pixel, and averaging the predicted values over each patch. 3D-VDP is able to capture both local and global attributes of experienced visual discomfort by modeling the mean firing rates of a variety of depth-sensitive cells in the middle temporal (MT) visual area. These cells have a wide range of receptive field shapes and sizes, hence are responsive to depth stimuli that occur over corresponding spatial scales [8].

As we will show later, using the predicted 3D-VDP scores as proxy ground-truth labels to pre-train the patch-based CNN which yields a performance improvement of more than 12%, as compared to using only raw disparity values as patch labels (Section IV).

D. DeepVDP Learning Framework

Fig. 3 depicts the overall DeepVDP training framework, which occurs in two consecutive steps: pre-training of the patch-based CNN by regressing onto the proxy ground-truth labels generated by 3D-VDP, followed by training (fine-tuning) the pre-trained model onto the MOSs. These automatically extracted local features are subsequently aggregated into global features predictive of the degree of visual discomfort

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2851670, IEEE Transactions on Image Processing

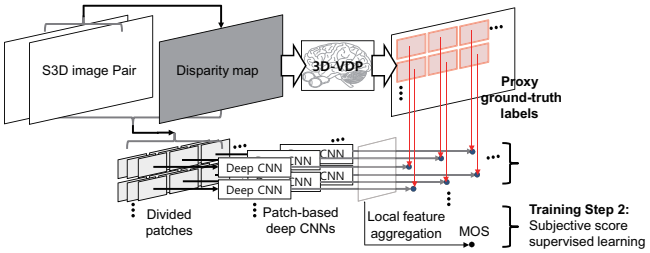


Fig. 3. Overall training framework of the DeepVDP model. Training step 1: the CNN model is locally regressed onto proxy ground-truth labels. Training step 2: the local features are aggregated and regressed onto MOS.

score experienced when viewing the S3D image. The specific two-step training processes are implemented as follows:

1) **Training step 1**: Proxy ground-truth patch labels are generated, and used to pre-train the patch-based CNN. Since the 3D-VDP score for each patch supplies only an approximation to ground-truth, the learning process requires fine tuning.
2) **Training step 2**: On each training image, the bundle of patch-based CNNs is trained onto the MOS, then the learned local features are aggregated into global features. The visual discomfort scores are predicted in a holistic manner by bi-directionally updating the global representations from the independently learned local features.

The contributions that we make are summarized as follows:

- The visual discomfort predictions delivered by the Deep-VDP model are highly correlated with subjective opinion scores, exceeding those of prior models. By extensive experimentation, we isolate and identify important factors that cause visual discomfort, including those related to human attention, depth distribution, and edge and texture orientations, amongst others. Our model is able to automatically extract features which cause visual discomfort.
- In order to overcome the lack of training data needed for deep learning, we deploy a high-performance existing VDP model to generate proxy ground-truth labels.
- A pre-training step using proxy ground-truth patch labels is developed and is shown to lead to a significant performance improvement.
- A process of local to global feature pooling is conducted during training, whereby the patch-based CNN is globally optimized to predict visual discomfort.

II. Proxy Ground-Truth Label Generation

A. Relevant Features Utilized by 3D-VDP

Vergence eye movements are controlled via feedback from vision to optomotor control. There are several cortical areas that are implicated in 3D visual perception, and there are numerous interconnections among them [28]. Recent studies have demonstrated that visual area MT plays a major role in disparity processing, and that disparity selectivity in this area is considerably stronger than in other cortical areas, including areas V1 and V4 [28]-[32]. Occurences of AVM are induced

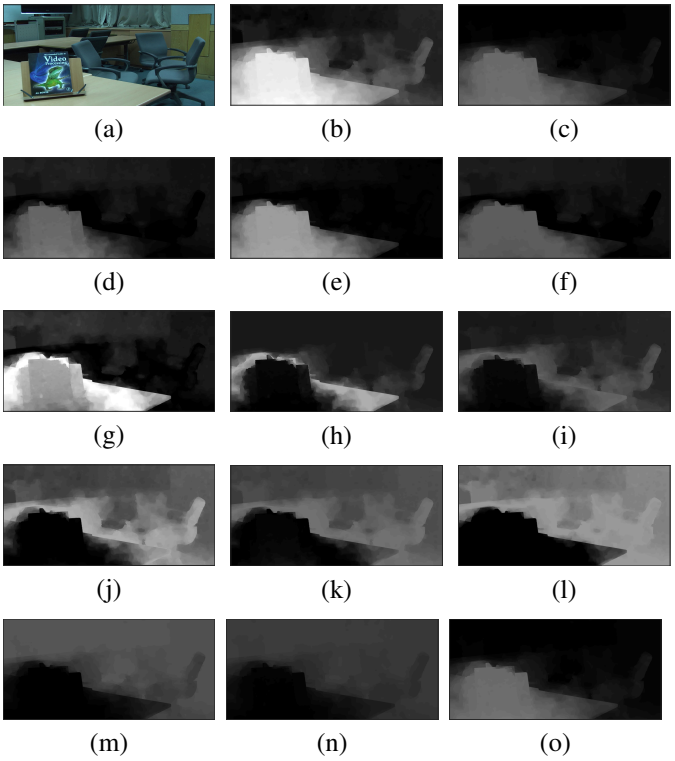


Fig. 4. Examples of estimated neural responses. (a) The left image of an S3D stereopair; (b) a computed disparity map by [31]; (c)-(o) neural response maps using 13 neuronal models of visual area MT.

in part by forced vergence eye movements, guided by neural activity transmitted from visual area MT. 3D-VDP models the tuning curves of area MT neurons as Gabor functions [30][33], using 13 typical tuning curves. In 3D-VDP, these 13 representative neurons are used to extract features, where the fitting parameters for the tuning functions are given in [30].

As an illustrative example, Fig. 4 (a) shows the left image of an S3D image pair, while Fig. 4 (b) shows a disparity map d_{map} computed on it, where brighter pixels represent higher disparities. Figs. 4 (c)–(o) depict estimated response maps using the 13 model tuning functions, where brighter regions represent disparity regions producing stronger responses. Further details of 3D-VDP can be found in [26].

B. Generating Proxy Ground-Truth labels

As depicted in Fig. 2, the generation of proxy ground-truth patch labels consists of the following three steps.

1) 3D-VDP model training: A support vector regressor (SVR) is employed to learn the 3D-VDP model $g_l(\cdot)$, which is trained on global features from entire images and MOSs as shown in Fig. 5 (a). Fig. 5 (a) also shows how the disparity map d_{map} is computed from a pair of S3D images, and how the neural responses are estimated using the aforementioned 13 tuning functions. In the 3D-VDP training step, thirteen feature maps f_1 to f_{13} are generated which, along with two absolute disparity features, are regressed onto the discomfort MOS using the SVR. In addition to the neuronal response maps, 3D-VDP deploys two additional disparity features f_{14} and f_{15}: the mean negative and positive disparities:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2851670, IEEE Transactions on Image Processing

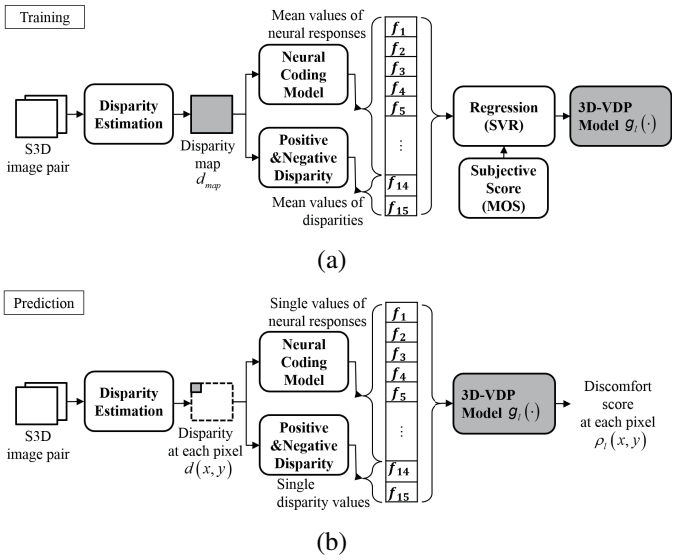


Fig. 5. Framework for obtaining the 3D-VDP score of each pixel. (a) Training: the 3D-VDP model $g_l(\cdot)$ is trained on 15 features, including neuronal response maps and 2 absolute disparity features. The model is regressed onto the MOS by using an SVR. (b) Prediction: the 15 local features are extracted at each coordinate (x, y) that are fed into the 3D-VDP model. The predicted proxy visual discomfort scores of the pixel are then used as proxy ground-truth labels for each patch.

$$f_{14} = \frac{1}{h_I \times w_I} \sum_x \sum_y d^-(x, y), \tag{1}$$

and

$$f_{15} = \frac{1}{h_I \times w_I} \sum_x \sum_y d^+(x, y), \tag{2}$$

where $d^-(x, y)$ and $d^+(x, y)$ are the negative and positive disparity maps and h_I and w_I are the number of horizontal and vertical pixels in each image, respectively.

2) Prediction of 3D-VDP score at each pixel: Visual discomfort scores are predicted by the trained model using the features extracted at each pixel, as shown in Fig. 5 (b). Let $\rho_l(x, y)$ be the 3D-VDP score at pixel coordinate (x, y), which is predicted using the trained model g_l:

$$\rho_l(x, y) = g_l\left(f_1(x, y), f_2(x, y), ..., f_{15}(x, y)\right), \tag{3}$$

where (x, y) are image coordinates and $f_1(x, y)$, $f_2(x, y)$, ..., $f_{15}(x, y)$ are features extracted at each pixel (x, y).

3) Assigning labels to patch: The proxy ground-truth label ρ_n^p of the n^{th} patch $\mathbf{p}_n$ is obtained by averaging the 3D-VDP scores over the patch:

$$\rho_n^p = \frac{1}{h_p \times w_p} \sum_{x,y \in \mathbf{p}_n} \rho_l(x, y) \tag{4}$$

where h_p and w_p are vertical and horizontal patch dimensions ($h_p = 16$ and $w_p = 18$). The reason for choosing this patch size is explained in Section IV.

Fig. 6. shows examples of the proxy ground-truth labels computed on three pairs of S3D images from the IEEE-SA database [21] (ISS7_0, ISS7_50, and ISS7_100). These

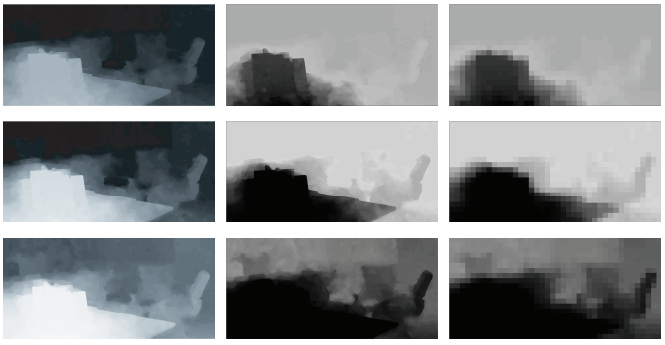


Fig. 6. Examples of obtained proxy ground-truth labels on three S3D image pairs, IEEE-SA database stereopairs (ISS7_0, ISS7_50 and ISS7_100) corresponding to each row. The first column contains the disparity maps computed using [31], the second column shows the 3D-VDP score at each pixel coded as an intensity map, and the last column represents the proxy ground-truth patch labels obtained by averaging the 3D-VDP scores of each patch.

three S3D image pairs were obtained from the same scene as Fig. 4 (a), but at different disparity ranges. The first, second, and third columns are disparity maps computed on them, the pixelwise 3D-VDP scores, and the proxy ground-truth patch labels, respectively. Brighter regions in the disparity maps indicate higher disparity values. The 3D-VDP score at each pixel was obtained using (3), where darker regions (lower 3D-VDP scores) might cause more severe visual discomfort. The proxy ground-truth patch labels are obtained using (4), and then they are used as labels during the DeepVDP model regression.

3D-VDP scores have been shown to correlate well with human visual discomfort judgments on the IEEE-SA database [17],[21]. For example, the normalized MOS values of ISS7_0, ISS7_50, and ISS7_100 are 0.48, 0.43, and 0.32, respectively. By comparison, the mean 3D-VDP discomfort predictions on the same images are 0.73, 0.59, and 0.25, respectively. Although the predicted 3D-VDP scores generally differ from the true MOS values, on average the MOS predictions are nicely linear and monotonic with, and accurately predict true MOS [17]. However, while errors are not unexpected, their occurrences motivate a second step of careful fine tuning on human scores.

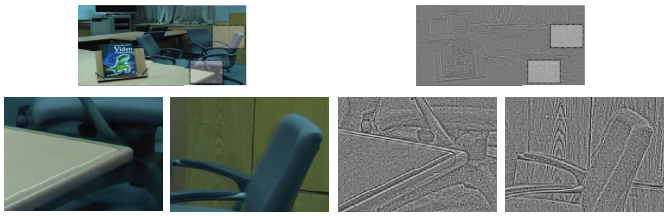
III. Deep Visual Discomfort Predictor

After generating proxy ground-truth patch scores using the 3D-VDP model, the DeepVDP model is then trained and tested on them. Next, we define the various inputs and explain how we train and test the DeepVDP model.

A. Input Generation

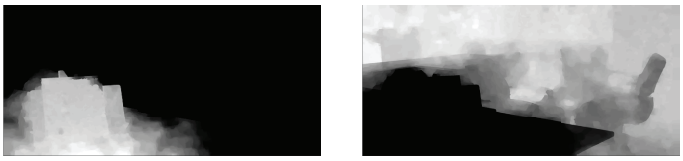
The input to the learning DeepVDP model consists of three channels: a normalized cyclopean image, and positive and negative disparity maps. Each input image is divided into non-overlapped patches of size $h_p \times w_p$. Let $\mathbf{p}_{cn}$, $\mathbf{p}_{dn}^-$ and $\mathbf{p}_{dn}^+$ be the n^{th} patches from the normalized cyclopean image $\hat{I}_c$, and negative and positive disparity maps d_{map}^- and d_{map}^+, respectively. These patches are spatially aligned. In this way, the three-channel patch set

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2851670, IEEE Transactions on Image Processing



(a) (b)

Fig. 7. (a) Synthesized cyclopean image of "ISS7_0" (upper image). (b) Normalized cyclopean image (upper image). The lower images are zoomed versions of the two images.



(a) (b)

Fig. 8. (a) Negative disparity map d^-_{map}, where brighter regions are in front of the screen. (b) Positive disparity map d^+_{map} where brighter regions are behind the screen.

$\mathbf{P} = \left\{\left\{\mathbf{p}_{c1}, \mathbf{p}^-_{d1}, \mathbf{p}^+_{d1}\right\} ..., \left\{\mathbf{p}_{cN}, \mathbf{p}^-_{dN}, \mathbf{p}^+_{dN}\right\}\right\}$ is composed, where N is the total number of three-channel patches. The set $\mathbf{P}$ becomes the training dataset for DeepVDP. Note that the n^{th} patch $\left\{\mathbf{p}_{cn}, \mathbf{p}^-_{dn}, \mathbf{p}^+_{dn}\right\}$ also corresponds to the proxy ground-truth label ρ^p_n in (4). The way of obtaining three input channels is described in the following.

1) Cyclopean Image: A single normalized "cyclopean image" is computed from each S3D pair and used as an input image when training or implementing DeepVDP. This mimics the perceptual process of fusing the images from the two eyes into a single *cyclopean image* [37]. We use the left-right disparity compensated weighted combination in [38] to synthesize the cyclopean image as:

$$I_c(x,y) = W_l(x,y)\cdot I_l(x,y) + W_r(x+d,y)\cdot I_r(x+d,y) \quad (5)$$

where I_l and I_r are the left and right images of an S3D pair, respectively, and W_l and W_r are corresponding weights applied to them. The weights are normalized Gabor filter responses [38]. Fig. 7 (a) shows the cyclopean image synthesized from "ISS7_0.''

The obtained cyclopean image is then subjected to de-biasing (local mean subtraction) and divisive normalization by local energy. These processes have relevance to both neuronal modeling and to established natural scene models [39]. Normalization can also induce improved performance of supervised learners [40] and can reduce training time [41]. These processes have also been used to define 2D and S3D picture quality predictors that are highly sensitive to deviations of natural image statistics caused by distortion [42]-[45]. The normalized cyclopean image is used as an input to DeepVDP, since the underlying statistics of image texture relates to stereopsis and the degree of experienced visual discomfort [46][47]. The processes of de-biasing and normalization on the cyclopean image I_c are given by:

$$\mu(x,y) = \sum_{k=-K}^{K}\sum_{l=-L}^{L} w_{k,l} I_c(x+k, y+l), \quad (6)$$

$$\sigma(x,y) = \sqrt{\sum_{k=-K}^{K}\sum_{l=-L}^{L} w_{k,l}\left[I_c(x+k,y+l) - \mu(x,y)\right]^2}, \quad (7)$$

where $w = \{w_{k,l} | k = -K, ..., L, l = -L, ..., L\}$ is a 2D circular symmetric Gaussian weighting function that is sampled out to 3 standard deviations ($K = L = 3$) and rescaled to unit volume [42]. The preprocessed cyclopean image is subsequently obtained as

$$\hat{I}_c(x,y) = \frac{I_c(x,y) - \mu(x,y)}{\sigma(x,y) + 1}, \quad (8)$$

which is identical to the preprocessing used in the BRISQUE and NIQE blind picture quality models [42][44].

Thus, a normalized image set $\hat{\mathbf{I}}_c = \left\{\hat{I}_{c1}, \hat{I}_{c2}, ..., \hat{I}_{cM}\right\}$ is generated, where M represents the total number of S3D image pairs in the training set. Fig. 7 (b) illustrates a normalized cyclopean image. The normalized images $\hat{\mathbf{I}}_c$ are the inputs to DeepVDP.

2) Positive and Negative Disparity Maps: The second and third inputs to DeepVDP are the negative and positive disparity maps d^-_{map} and d^+_{map}, which are extracted from the computed disparity map d_{map}:

$$\begin{aligned} d^-_{\text{map}}(x,y) &= \min\left(0, d_{\text{map}}(x,y)\right) \\ d^+_{\text{map}}(x,y) &= \max\left(0, d_{\text{map}}(x,y)\right). \end{aligned} \quad (9)$$

A variety of studies have been directed towards understanding the relationships between positive and negative disparities and their different effects on experienced visual discomfort [7][13][8][48]. While negative disparities are often regarded as having a greater contribution to feelings of visual discomfort than positive disparities, this is not always the case, hence we model positive and negative disparities separately. Thus, distinct maps d^-_{map} and d^+_{map} are computed to account for the relative distributions of negative disparities (implied depths in front of the viewing screen) and positive disparities (implied depths behind the viewing screen). Figs. 8 (a) and 8 (b) show examples of d^-_{map} and d^+_{map}, respectively. Darker regions indicate lower disparity values that are closer to zero disparity (screen depth).

B. Patch-based CNN Architecture

Fig. 9 illustrates the overall architecture of the patch-based CNN. The proposed model consists of two convolutional layers, three consecutive fully connected layers, and a fully connected regression layer at the end. In the first and second convolutional layers, a 5×5 filter is used without pooling, and 48 and 64 kernels are used, respectively. Let $g_\theta(\cdot)$ be the feature vector extractor ($\mathbb{R}^{3\times h_p \times w_p} \rightarrow \mathbb{R}^{200}$), which includes two convolutional layers and three consecutive fully connected layers parameterized by θ. In particular, the 200 feature dimensions are heuristically determined based on the performance as a function of feature dimension. Let $g_{\phi_1}(\cdot)$ be

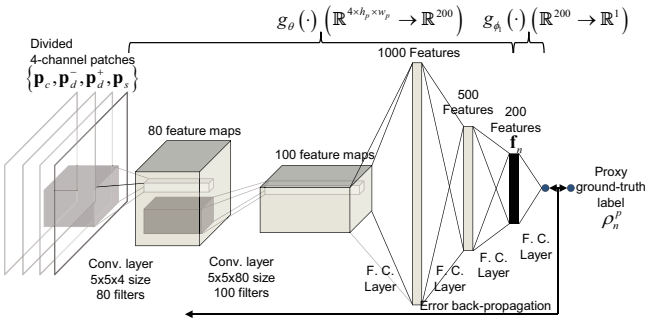


Fig. 9. CNN-based proxy ground-truth label training. The 200-dimensional local feature vector $\mathbf{f}_n$ is extracted from the last hidden node (black).

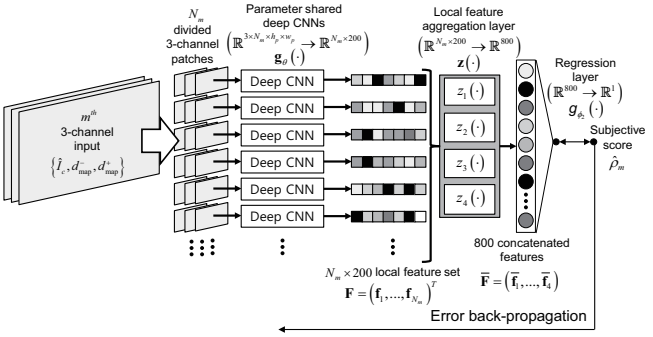


Fig. 10. Framework of training step 2. The local features $\mathbf{F}$ are extracted independently through the shared CNNs and pooled into global features $\bar{\mathbf{F}}$ at the local feature aggregation layer $\mathbf{z}(\cdot)$. The model parameters are then iteratively trained using the MOS.

the regression function ($\mathbb{R}^{200} \rightarrow \mathbb{R}^1$) parameterized by ϕ_1 at the end of the network model, as illustrated in Fig. 9. Thus, the 200-dimensional feature vector $\mathbf{f}_n = (f_{n,1}, f_{n,2}, ..., f_{n,200})$ is extracted by $g_\theta\left(\left\{\mathbf{p}_{cn}, \mathbf{p}^+_{dn}, \mathbf{p}^-_{dn}\right\}\right)$ from the n^{th} patch. In each layer, a leaky rectified linear unit (LReLU) is applied as a nonlinear activation unit with a small constant $\alpha = 0.1$ [50]. This has advantages for training while simplifying the back-propagation process, enhancing optimization, and preventing saturation owing to differentiation [51].

C. Training Step 1: Proxy Ground-Truth Patch Label Regression

In the first training step, the patch-based CNN is trained onto the proxy ground-truth label to learn local abstractions of the data. Each patch $\left\{\mathbf{p}_{cn}, \mathbf{p}^-_{dn}, \mathbf{p}^+_{dn}\right\}$ is an input to the model, the proxy ground-truth label for the n^{th} patch ρ^p_n is an output to the model, and the network parameters Θ_1 are optimized to minimize the loss:

$$\Theta^*_1 = \arg\min_{\Theta_1} \ell_1\left(\left\{\mathbf{p}_{cn}, \mathbf{p}^-_{dn}, \mathbf{p}^+_{dn}\right\}, \rho^p_n; \Theta_1\right), \tag{10}$$

where the loss function $\ell_1(\cdot)$ indicates the output values computed by the 3-channel input patch $(\mathbf{p}_{cn}, \mathbf{p}^-_{dn}, \mathbf{p}^+_{dn})$ and the two-stage feedforward network $(g_{\phi_1}(\cdot), g_\theta(\cdot))$, where the loss is determined by the mean squared error (MSE) between the network output and the proxy ground-truth patch labels:

$$\begin{aligned}&\ell_1\left(\left\{\mathbf{p}_{cn}, \mathbf{p}^-_{dn}, \mathbf{p}^+_{dn}\right\}, \rho^p_n; \Theta_1 = (\theta, \phi_1)\right) \\ &= \frac{1}{N_T}\sum_{n=1}^{N_T}\left(g_{\phi_1}\left(g_\theta\left(\left\{\mathbf{p}_{cn}, \mathbf{p}^-_{dn}, \mathbf{p}^+_{dn}\right\}\right)\right) - \rho^p_n\right)^2,\end{aligned} \tag{11}$$

where N_T is the size of the training set. In order to optimize the model parameters, stochastic gradient descent is employed in the back-propagation process over a mini-batch $B = 500$, i.e., $\frac{1}{B}\sum_{n=1}^{B}\frac{\partial \ell_1\left(\left\{\mathbf{p}_{cn}, \mathbf{p}^-_{dn}, \mathbf{p}^+_{dn}\right\}, \rho^p_n; \Theta_1\right)}{\partial \Theta_1}$. The initialization and optimization techniques are adopted from [52][53]. In order to prevent falling into local minima and overfitting, batch normalization is used in training except g_{ϕ_1}, which can be achieved through reduction of the internal covariate shift problem [54].

D. Training Step 2: Subjective Score Regression

Each locally trained patch-based CNN model only captures limited information descriptive of visual discomfort, since it is supervised using the proxy ground-truth patch labels. Therefore, the learned local features are updated to global features representing holistic visual discomfort, following the last hidden node (black) of each model in Fig. 9. In training step 2, the model is trained again with the grouped patches obtained from each input image, as illustrated in Fig. 10. Further, the local feature aggregation layer $\mathbf{z}(\cdot)$ is embedded into the learning process. This is combined with the process of optimizing the model parameters in the end-to-end framework.

TABLE I
A LIST OF VARIABLES REPRESENTING FEATURES

Variables	Meaning
$f_{n,l}$	feature value of local patch
$\bar{f}_{k,l}$	pooled feature value of $f_{n,l}$
$\bar{\mathbf{f}}_k$	aggregated feature vector of an image
$\bar{\mathbf{F}}$	concatenated feature vector of $\bar{\mathbf{f}}_k$

Let N_m be the number of divided patches in the m^{th} S3D image. A set of feature vectors $\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_{N_m})^T$ is obtained from a single S3D image by $\mathbf{g}_\theta\left(\left\{\hat{I}_{cm}, d^-_{\text{map},m}, d^+_{\text{map},m}\right\}\right) = \left\{g_\theta\left(\left\{\mathbf{p}_{c1}, \mathbf{p}^-_{d1}, \mathbf{p}^+_{d1}\right\}\right), ..., g_\theta\left(\left\{\mathbf{p}_{cN_m}, \mathbf{p}^-_{dN_m}, \mathbf{p}^+_{dN_m}\right\}\right)\right\}$ ($\mathbb{R}^{3\times N_m \times h_p \times w_p} \rightarrow \mathbb{R}^{N_m \times 200}$), where $\mathbf{g}_\theta$ represents the bundle of patch-based CNNs g_θ.

The feature aggregation layer $\mathbf{z} = \{z_1(\cdot), z_2(\cdot), z_3(\cdot), z_4(\cdot)\}$ consists of four pooling functions ($\mathbb{R}^{N_m} \rightarrow \mathbb{R}^4$). Let $\bar{\mathbf{F}} = \left(\bar{\mathbf{f}}_k | k = 1, 2, 3, 4\right)$ be the global feature vector obtained from each pooling function, where $\bar{\mathbf{f}}_k = (\bar{f}_{k,l} | l = 1, 2, ..., 200)$. The local features of the n^{th} patch $f_{n,l}$ are then pooled into $\bar{f}_{k,l}$:

$$\bar{f}_{1,l} = z_1(\mathbf{F}) = \frac{1}{N_m}\sum_{n=1}^{N_m} f_{n,l}, \tag{12}$$

$$\bar{f}_{2,l} = z_2(\mathbf{F}) = \frac{1}{N_m}\sum_{n=1}^{N_m}\left(f_{n,l} - \bar{f}_{1,l}\right)^2 \tag{13}$$

$$\bar{f}_{3,l} = z_3(\mathbf{F}) = \frac{1}{N^p_m}\sum_{n>n^{p+}} f^h_{n,l}, \tag{14}$$

$$\bar{f}_{4,l} = z_4(\mathbf{F}) = \frac{1}{N^p_m}\sum_{n<n^{p-}} f^h_{n,l} \tag{15}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2851670, IEEE Transactions on Image Processing

where $l = 1, ..., 200$ indexes the features from g_θ. As a result, each $\bar{\mathbf{f}}_k$ has 200 dimensions. Here, n^{p+} and n^{p-} indicate the upper and lower p^{th} percentiles in the histogram of local features $f^h_{n,l}$, respectively. N^p_m represents the number of p-percentile proxy visual discomfort scores, i.e., $N^p_m = N_m \cdot p/100$. In our implementation, we set $p = 10$ [57].

Consequently, the concatenated form of the global feature vector $\bar{\mathbf{F}} = (\bar{\mathbf{f}}_1, ..., \bar{\mathbf{f}}_4)$ is obtained as $\mathbf{z}(\cdot)$ ($\mathbb{R}^{N_m \times 200} \rightarrow \mathbb{R}^{800}$). A list of the relevant variables is given in Table I. Subsequently, $\mathbf{g}_\theta$ for the m^{th} input image is regressed onto its corresponding MOS $\hat{\rho}_m$ by minimizing the objective function:

$$\Theta_2^* = \arg\min_{\Theta_2} \ell_2 \left(\left\{ \hat{I}_{cm}, d^-_{\text{map},m}, d^+_{\text{map},m} \right\}, \hat{\rho}_m; \Theta_2 \right) \quad (16)$$

where the loss function $\ell_2(\cdot)$ in training step 2 is the MSE between the predicted S3D visual discomfort level and the MOS of the M_T training input images:

$$\ell_2 \left(\left\{ \hat{I}_{cm}, d^-_{\text{map},m}, d^+_{\text{map},m} \right\}, \hat{\rho}_m; \Theta_2 = (\theta, \phi_2) \right) = \quad (17)$$
$$\frac{1}{M_T} \sum_{m=1}^{M_T} \left(g_{\phi_2} \left(\mathbf{z} \left(\mathbf{g}_\theta \left(\left\{ \hat{I}_{cm}, d^-_{\text{map},m}, d^+_{\text{map},m} \right\} \right) \right) \right) - \hat{\rho}_m \right)^2,$$

where $g_{\phi_2}(\cdot)$ is the regression function at the end of the model structure with parameter ϕ_2, which predicts the visual discomfort score of the S3D image ($\mathbb{R}^{800} \rightarrow \mathbb{R}^1$), as illustrated in Fig. 10.

An essential consideration is the size of the mini-batch deployed in the stochastic gradient descent, because the proposed model is a patch-based structure. Unlike training step 1, training step 2 requires an image-based structure. Rather than training the model over the entire image domain, to achieve reliable local feature aggregation, the size of the mini-batch is set to be identical to the number of patches which comprise the m^{th} S3D image, i.e., $B = N_m$. The other settings for the model training are the same as in training step 1.

IV. Experimental Results

A. Performance of DeepVDP

1) Dataset: To verify the performance of DeepVDP, the IEEE-SA S3D image database was used [21]. IEEE-SA consists of 800 S3D image pairs obtained using a built-in twin-lens camera system. The size of each image is 1920×1080. The database is organized into eight categories encompassing a diversity of shapes and depths, which are reasonably representative and challenging. The IEEE-SA stereo image database is composed of 160 convergence-sampled sets (i.e., five S3D image pairs with different disparity ranges for each set) such that each content category contains 20 sets. The size of the LCD monitor used to display S3D images to subjects was 46 inches, while the design of the experimental environment followed the standard recommendation ITU-R BT. 500-11 [22]. The MOS value of each S3D image pair falls within the range $[1, 5]$, where a lower value represents higher visual discomfort. Before training, we normalized the MOS to the range $[-1, 1]$.

To validate the generality of the model, we also used the IVY LAB S3D image database [23], which is composed of 120 real scenes captured using a 3D digital camera with dual lenses. The IVY LAB database consists of various categories of images of resolution 1920×1080 pixels, including both indoor and outdoor scenes containing various objects (humans, trees, buildings, man-made objects, etc.). The size of the LCD monitor used in the IVY study was 40 inches, while the experimental design also followed ITU-R BT. 500-11 [22]. The MOS value of each IVY S3D image pair falls in the range $[1, 5]$, which we also normalized to $[-1, 1]$.

TABLE II
LCC and SROCC comparison on the IEEE-SA VDP database.

VDP model	LCC	SROCC	RMSE
Yano *et al.* [10]	0.4030	0.3361	0.7654
Nojiri *et al.* [13]	0.6938	0.6063	0.5923
Choi *et al.* [14]	0.6732	0.5866	0.6102
Kim *et al.* [15]	0.7040	0.6172	0.5579
Park *et al.* [17]	0.8524	0.7785	0.4247
Park *et al.* [8]	0.8505	0.7784	0.4234
Oh *et al.* [7]	0.8590	0.7887	0.4108
DeepVDP	**0.8849**	**0.8164**	**0.3631**

TABLE III
LCC and SROCC comparison on the IVY LAB VDP database.

VDP model	LCC	SROCC	RMSE
Yano *et al.* [10]	0.4105	0.3458	0.7414
Nojiri *et al.* [13]	0.7025	0.6127	0.5897
Choi *et al.* [14]	0.6821	0.5977	0.5922
Kim *et al.* [15]	0.7110	0.6246	0.5313
Park *et al.* [17]	0.8623	0.7813	0.4121
Park *et al.* [8]	0.8614	0.7872	0.4130
Oh *et al.* [7]	0.8653	0.7925	0.4052
DeepVDP	**0.8885**	**0.8254**	**0.3612**

2) Performance Measurement: Well-known performance measurements were utilized to benchmark DeepVDP against the performances of previous S3D VDP models: the Pearson linear correlation coefficient (LCC), Spearman rank-order correlation coefficient (SROCC) and root mean square error (RMSE). The LCC and SROCC measure the degree of linearity and monotonicity between the predicted discomfort scores and the ground-truth scores. The RMSE measures the accuracy (average distance) between the predicted scores and the MOS. The same 80% of the IEEE-SA and IVY LAB databases were randomly selected for both training steps 1 and 2, while the remaining 20% of the dataset was used for testing. Each correlation coefficient is an average of values obtained over 50 training and testing iterations. To increase the size of the training set in step 2, we swapped the left and right images after left-right flipping each S3D image. This process maintains the same depth distribution as the original stereopair, where it is reasonably assumed that the flipped stereopair is associated with the same MOS as the original stereopair. We compared the predictions of the proposed model against those of prior seven VDP models developed by Yano *et al.* [10], Nojiri *et al.* [13], Choi *et al.* [14], Kim *et al.* [15], Park *et al.* [17], Park *et al.* [8], Oh *et al.* [7]. Tables II and III compare the LCC and SROCC of these VDP models on the IEEE-SA and IVY LAB databases, respectively. Clearly, DeepVDP delivers improved predictive performance compared to the other models in terms of both linear accuracy and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2851670, IEEE Transactions on Image Processing

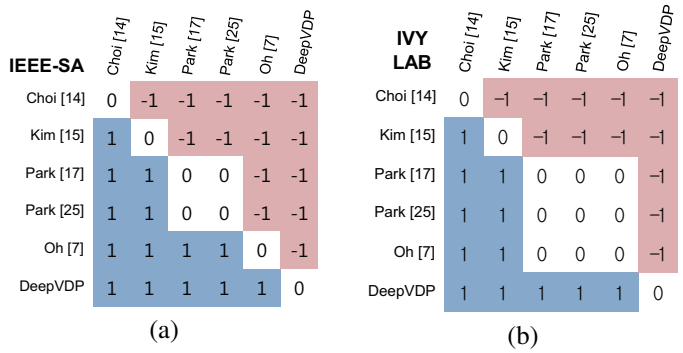


Fig. 11. Results of the t-test on the LCC values from the (a) IEEE-SA and (b) IVY LAB VDP databases.

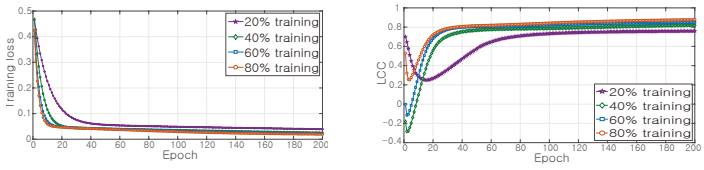


Fig. 12. Variation of training loss in (a), and LCC in (b) for each percentage of the training set as a function of the number of epochs in step 2.

monotonicity.

Also, we conducted a t-test simulation on the LCC values (over 50 trials) of all pairs of prediction models. The t-test is performed in this context to determine whether the superiority of the performance of one model over another is statistically significant. The results of the t-tests on the IEEE-SA and IVY LAB databases are illustrated in Fig. 11. The symbols "1", "0" and "-1" indicate that the performance of the model in the row is statistically better, indistinguishable or worse, respectively, than the competitor in the column. We fixed the significance level at p=5% in the comparisons. However, when there were multiple comparisons involving more than three models, to decrease type I errors, we applied Bonferroni correction [58], yielding a significance level $p/\binom{6}{2} = 0.417\%$. On the IEEE-SA and IVY LAB VDP databases, the DeepVDP model provided to be statistically better than all of the other VDP models.

Additionally, to study the dependency of the performance of DeepVDP on the relative proportion of samples between the training and test sets, we measured the training loss and LCC when training on four different percentages of the overall dataset (20%, 40%, 60%, and 80%) on the IEEE-SA database. Furthermore, we also observed these values against the number of epochs (1–200) to verify whether the training loss properly converged without the gradient diverging, and to analyze the trend of the performance improvement. Figs. 12 (a) and 12 (b) show the variation of training loss and LCC as a function of the percentage of the dataset that was trained on. The training loss rapidly decreased during the early iterations, which demonstrates that the initial values of the model parameters were reliably being optimized onto the MOS. When 80% of the dataset was used for model training, the training loss rapidly converged until the 15^{th} epoch. The accuracy then stabilized to a correlation of approximately 88% against MOS around the 180^{th} epoch, as shown in Fig. 12 (b). When 60% of the dataset was trained on, the convergence was slightly slower. During the early stages, the LCC score obtained using 60% of the dataset for training was higher than when 80% of the dataset was used for training; however, the performance of the model improved very little after 40 epochs. Nevertheless, the performance of the model was close to that achieved when 80% of the dataset was used for training, attaining an accuracy of approximately 84%. When 40% of the dataset was used to train the model, the model parameters slowly converged as compared to when 80% of the dataset was used for training, and the final achieved correlation was again slightly lower at 82%. Finally, when only 20% of the subset was used for training, the accuracy was significantly lower than in the other cases.

It is also important to discuss the relationship that exists between expressed visual discomfort and the parameters defining the viewing conditions (display size, resolution, viewing distance, and so on). It is known that the degree to which visual discomfort is experienced depends on the viewing parameters [60]. Because of this, it is inadvisable to extrapolate the performance of DeepVDP to arbitrary viewing conditions, since it was trained on human subjective data taken under a specific viewing environment. Still, the experimental MOS were obtained under typical viewing conditions, and as we shall see, the trained model performs well on other databases having somewhat different viewing conditions, allowing us to impute a degree of robustness to the DeepVDP model.

3) Cross Dataset Test: We also conducted an additional cross-database evaluation to validate the generality of the model. Towards this end, we utilized the EPFL S3D image database [61]. After learning a discomfort prediction model using 80% of the training data from the IEEE-SA and IVY LAB databases, predicted visual discomfort scores were inferred on the S3D images in the EPFL database using the trained model parameters. Tables IV and V tabulates the LCC and SROCC of the VDP models following this train-test sequence. As shown

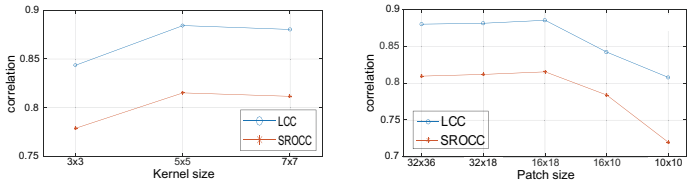
TABLE IV
LCC AND SROCC COMPARISON ON THE EPFL DATABASE, WHERE THE PREDICTION MODEL WAS TRAINED USING THE IEEE-SA DATABASE.

VDP model	LCC	SROCC	RMSE
Choi *et al.* [14]	0.7731	0.8231	0.4539
Kim and Sohn [15]	0.8693	0.8720	0.4296
Park *et al.* [17]	0.8893	0.8902	0.3925
Park *et al.* [8]	0.8882	0.8896	0.3892
Oh *et al.* [7]	0.8864	0.8812	0.3934
DeepVDP	**0.9102**	**0.9154**	**0.3613**

TABLE V
LCC AND SROCC COMPARISON ON THE EPFL DATABASE, WHERE THE PREDICTION MODEL WAS TRAINED USING THE IVY LAB DATABASE.

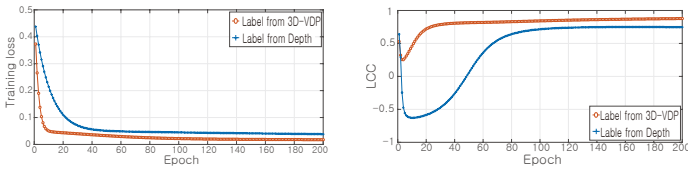
VDP model	LCC	SROCC	RMSE
Choi *et al.* [14]	0.7528	0.8012	0.4721
Kim and Sohn [15]	0.8471	0.8503	0.4378
Park *et al.* [17]	0.8562	0.8617	0.4202
Park *et al.* [8]	0.8530	0.8609	0.4198
Oh *et al.* [7]	0.8591	0.8631	0.4177
DeepVDP	**0.8782**	**0.8876**	**0.3934**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2851670, IEEE Transactions on Image Processing



(a) (b)

Fig. 13. Plots of LCC and SROCC as functions of (a) kernel size and (b) patch size.



(a) (b)

Fig. 14. Plots of (a) training loss, and (b) LCC as functions of epoch number, using labels from 3D-VDP (red) and from positive/negative depths (blue).

in Tables IV and V, the overall performance of our model was significantly better than that of the other VDP algorithms.

B. Analysis of DeepVDP

1) Effects of Kernel and Patch Sizes: In order to analyze the effects of the kernel and patch sizes on model performances, three different kernel sizes (3×3, 5×5, and 7×7) and five different patch sizes (32×36, 32×18, 16×18, 16×10, and 10×10) were used. The kernel size of the convolutional layer was fixed to 5×5 when the patch size was varied; thus, only the sizes of the extracted feature maps from the first and second convolutional layers were varied in the experiments. Furthermore, when the kernel size was varied, the images were divided into patches of fixed size 16×18 without overlap.

As may be seen in Fig. 13 (a), the correlation scores were almost the same for kernel sizes of 5×5 and 7×7. The 5×5 kernel, which captures a bit more detail, gave slightly improved performance [62]. However, the small 3×3 kernel was unable to adequately capture spatial relationships between neighboring pixels when seeking to learn meaningful low-level features in the convolutional layers [63]. As shown in Fig. 13 (b), almost the same performances were achieved when using 32×36, 32×18, and 16×18 patches. However, on small patches such as 16×10 and 10×10, the experimental results indicate that the extracted features did not as adequately represent visual discomfort.

2) Effects of Proxy Ground-Truth Labels: In DeepVDP, the proxy ground-truth labels (3D-VDP scores) were used to supervise training step 1. To study the dependency of the model on the assigned proxy ground-truth labels, we applied different proxy ground-truth patch labels defined using only two features in (3), the negative and positive disparities f_{14} and f_{15}. Hence $g_l(\cdot)$ was also trained onto MOS using the negative and positive disparities computed from the S3D images, thereby generating the proxy visual discomfort scores. 3D-VDP scores were not used in these processes. Subsequently,

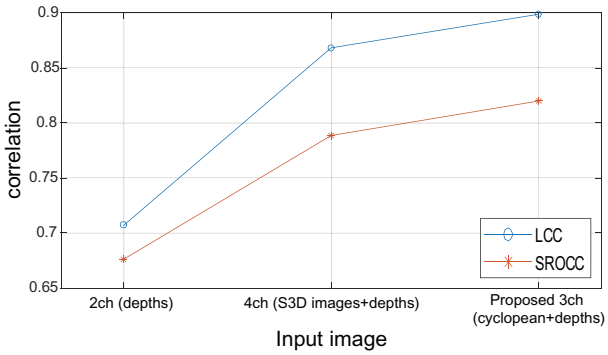
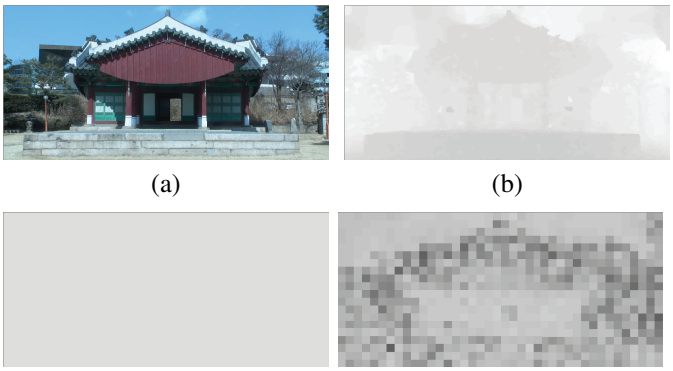


Fig. 15. Comparison of LCC and SROCC according to channel of input images.



(a) (b) (c) (d)

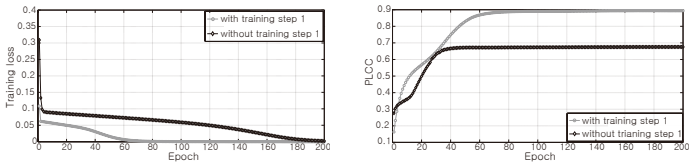
Fig. 16. Failed example using only two input channels (positive and negative disparities) to train DeepVDP. (a) Left image of S3D image pair "OSL6_75". (b) Computed disparity map. (c) Proxy ground-truth labels from 3D-VDP scores. (d) Predicted visual discomfort score map.

the ground-truth label of each patch was obtained by (4), and the patch-based CNN model was trained onto it.

Figs. 14 (a) and 14 (b) plot the training loss and LCC for these choices of proxy ground-truth patch labels. As shown in Fig. 14 (a), when the proxy ground-truth patch labels were assigned using only depth information, the training loss decreased more slowly as compared to using the proxy ground-truth labels assigned by 3D-VDP. After approximately 60 epochs, the training loss in the former case was even higher than in the latter case, although it nearly converged. As shown in Fig. 14 (b), the predictive performance when using proxy ground-truth patch labels from depth only was significantly reduced (the LCC was approximately 0.74) than when the proxy ground-truth labels generated using 3D-VDP were used. The result strongly suggests that the quality of the proxy ground-truth patch labels significantly affects how well representative local features were learned in training step 1. The proposed proxy ground-truth patch labeling using 3D-VDP scores appears to have effectively enhanced the prediction of VDP scores in this context.

3) Effects of Input Images: In order to investigate the efficacy of the selected input images, we conducted additional experiments using other input images. Initially, two input images composed of only negative and positive disparities were used (two channels). As shown in Fig. 15, in this

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2851670, IEEE Transactions on Image Processing



(a) (b)

Fig. 17. (a) Variation of training loss and (b) variation in PLCC for each percentage of the training set according to the number of epochs without using step 1.

case, the LCC and SROCC significantly decreased to $\sim 71\%$ and $\sim 68\%$, respectively. Clearly, information was lost when applying only depth information, without texture. In Fig. 16, the failed proxy visual discomfort prediction on exemplar "OSL6_75" strongly suggests that using only the two channel inputs (negative and positive disparities) is insufficient to train DeepVDP. As shown in Fig. 16 (c), the proxy ground-truth labels on this image are uniformly distributed with high values; the overall image is regarded as comfortable by 3D-VDP. However, the predicted score map in Fig. 16 (d) is quite noisy. Moreover, the recorded MOS of this image is 4.38, one of the most comfortable S3D scenes in the IEEE-SA database. Clearly, disparity in isolation is an inadequate input to train DeepVDP to successfully predict experienced visual discomfort.

Next, the prediction model was trained using four channels, consisting of the left/right S3D images along with the negative/positive disparities. The cyclopean image was still excluded. In this case, the correlation scores improved as compared to the two-channel case; however, the performances were still inferior (LCC and SROCC of approximately 0.86 and 0.78, respectively). One may reason that the cyclopean image formed by the process of binocular fusion is implicated in the experience of visual discomfort.

4) Advantage of Two-Stage Training Steps: To study the performance improvement gained by using the two-stage training procedure, we trained the model using only training step 2 (MOS supervised learning), without training step 1. Figs. 17 (a) and (b) show the results of training loss and LCC when training step 1 is omitted. As shown in Fig. 17 (a), the training loss significantly lags when training step 1 is not used. Fig. 17 (b) shows that the prediction performance was significantly degraded when the model parameters were learned only using training step 2, and that the difference was sustained. These results show that highly representative local features are learned in training step 1.

5) Visualization of Predicted proxy Visual Discomfort Scores: Following training step 1, the DeepVDP model is able to predict proxy visual discomfort scores based on proxy ground-truth label supervision. In Fig. 18, the predicted proxy visual discomfort score maps and their ground-truths are compared. Each row of Fig. 18 shows results from each image category in the IEEE-SA database: ISS, ISL, INS, INL, OSS, OSL, ONS, and ONL, respectively. The leftmost column of Fig. 18 shows the left images of the S3D image pairs, the center column shows the ground-truth label maps obtained from 3D-VDP scores, while the rightmost column shows the predicted proxy visual discomfort patch scores obtained from the learned model, respectively. It is evident that the learned model predicts the proxy visual discomfort scores consistently, which augments VDP performance in training step 2.

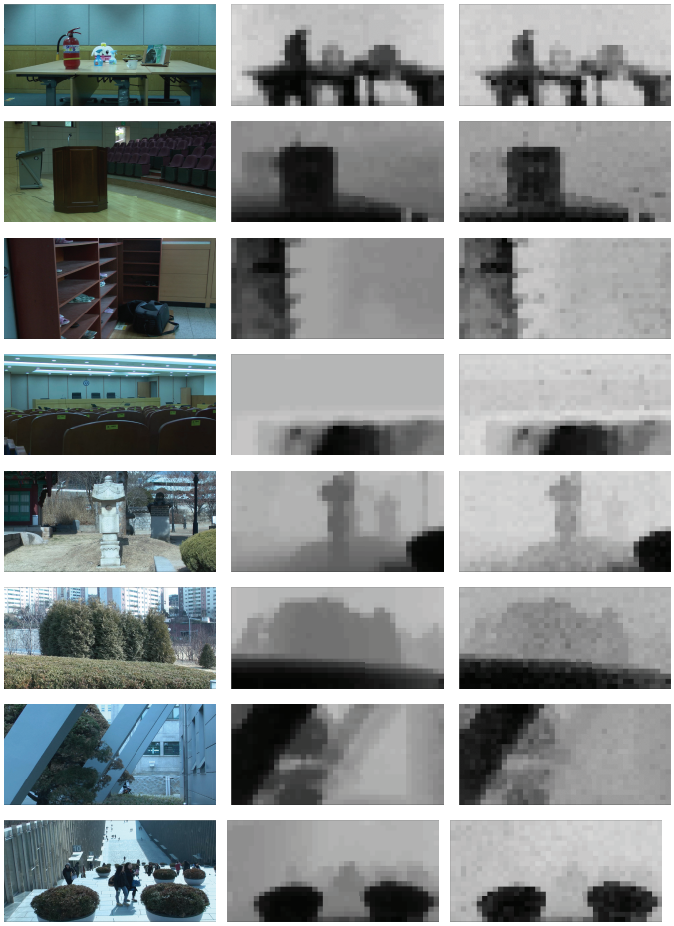


Fig. 18. Examples of predicted **proxy** visual discomfort **patch** score maps using training step 1, for each image category in the IEEE-SA database: the rows correspond to categories ISS, ISL, INS, INL, OSS, OSL, ONS and ONL, respectively. The left column shows the left images of the S3D image pairs, the center column shows the proxy ground-truth label maps generated using 3D-VDP, and the rightmost column shows the predicted **proxy** visual discomfort score **patch** maps, respectively.

6) Feature Map Visualization: Although the proxy ground-truth patch labels are constructed using only disparity information in training step 1, texture information is also involved in the prediction of visual discomfort by DeepVDP. In order to analyze the kinds of features being extracted by the DeepVDP kernels in detail, we captured the feature maps from the first convolutional layer. As was stated in Section III-A, the input images are divided into patches of size 16×18 without overlap. These are fed into the patch-based CNN, where the feed-forward value becomes the predicted visual discomfort score. Since the kernel size of the first convolutional layer was 5×5 without pooling, padding, or striding, the output size of each patch was 12×14. Each input image is comprised of $N_m = 40 \times 20$ patches. The 480×280 feature map can be obtained by gathering the output of each patch. Since the number of kernels in the first convolutional layer is 48, then 48 feature maps are observable. Fig. 19 shows six selected

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2851670, IEEE Transactions on Image Processing

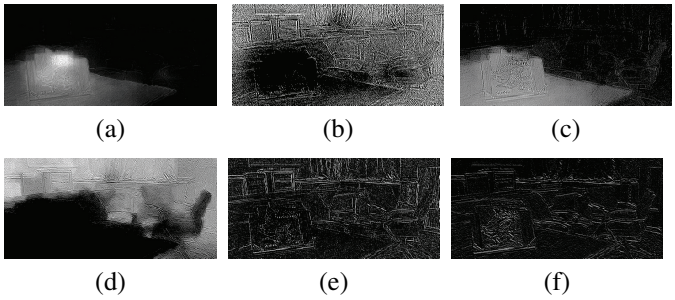


Fig. 19. Six examples of the feature maps obtained by the first convolutional layer on input S3D images "ISS7_0" depicted in Figs. 7 and 8. Brighter regions represent highly activated regions.

examplar feature maps computed on the image "ISS7_0," where brighter regions represent highly activated regions.

- **Observation 1 (human attention)**: Fig. 19 (a) shows that salient regions are dominantly activated, indicating that some of the kernels heavily depend on human attention to extract meaningful local VDP features. However, a few kernels focus on non-salient regions, as shown in Fig. 19 (b). Although it is difficult to exactly quantify the importance of each feature map in the CNN structure, relatively more kernels were activated values in salient regions than in non-salient regions. This suggests that many discomfort predictive features relate to human attention. Notably, DeepVDP extracts not only local low-level features related to salient and non-salient regions, but also high-level features combining them.
- **Observation 2 (depth levels)**: Figs. 19 (c) and 19 (d) show the different feature maps, which indicate that the kernels selectively respond to the depth levels. Further, Fig. 19 (c) also shows that negative disparity regions are mainly activated, while Fig. 19 (d) shows the opposite. One may conclude that the parameters of DeepVDP are reliably optimized to predict visual discomfort, since the main cause of visual discomfort is AVM arising from forced 3D depths, as observed in Section I.
- **Observation 3 (edge and texture orientations)**: Fig. 19 (e) shows the feature map, which is activated at strong edges. As suggested in [66], luminance gradients related to transitions of visual attention are implicated in visual discomfort i.e., local features responsive to edges are useful abstractions for predicting visual discomfort. Further, Fig. 19 (f) shows that the corresponding kernel activates only certain directional texture components. This suggests that the orientation of texture also affects the level of experienced visual discomfort.

The six feature maps in Fig. 19 were selected as simple examples to visualize the learned local feature characteristics in terms of lower- and mid-level information. The other feature maps appear more complex; for example, one of them activates high frequencies in salient regions with negative depths. These might capture higher-level information, although they are extracted from the first convolutional layer. Therefore, it is expected that the parameters of later layers extract higher-level features, which embody more complicated VDP representations.

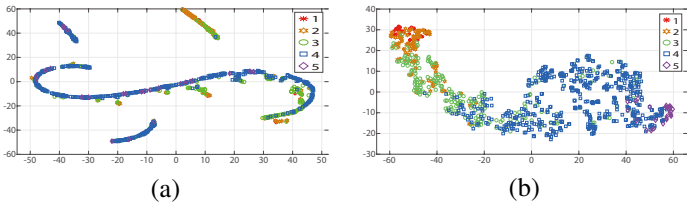


Fig. 20. Visualization of two-dimensional manifold obtained by t-SNE. Each point indicates an S3D image pair, and the points are labeled according to subjective opinions. Each graph represents (a) A manifold visualization of input features, (b) A manifold visualization of the aggregated feature vector.

7) Local Neighborhood Embedding: In order to demonstrate the method of extracting the feature vector of the learned DeepVDP model, the input features (i.e., four-channel input images) and the aggregated feature vector ($\mathbb{R}^{800}$) prior to the MOS regression layer g_{ϕ_2} are visualized. A well-known embedding algorithm, t-SNE [68] was employed as a way of reducing the high-dimensional data into a lower dimension. Fig. 20 shows a generated two-dimensional manifold obtained using t-SNE, where each point indicates an S3D image pair, and the points are labeled according to subjective opinions. Further, lower MOS represents more uncomfortable S3D image pairs. As shown in Fig. 20 (a), the input features in the reduced dimension are uncorrelated with MOS, and each S3D image pair is incoherently clustered. This indicates that the input features are insufficient to predict the level of visual discomfort. On the other hand, as shown in Fig. 20 (b), the data points in the graphed manifold can be clearly separated according to their MOS values when the extracted global features are used in t-SNE. The upper left part of the figure indicates severely uncomfortable S3D pairs, while the lower right part includes more comfortable scenes. This indicates that DeepVDP is properly trained onto MOS, and is effective for extracting meaningful features for VDP.

V. Conclusion

We explored a deep learning approach to the problem of predicting the degree of visual discomfort experienced when viewing S3D images. We showed that proxy ground-truth patch labels obtained using 3D-VDP adequately capture lower level representations of experienced visual discomfort to drive patch-based training of a CNN to accurately predict experienced visual discomfort. The results suggest that a network pretrained on proxy patch labels generated by the 3D-VDP model causes large amounts of latent variables to converge to more proper optima. Consequently, the proposed method outperforms previous VDP algorithms, demonstrating the potential of deep learning for S3D discomfort prediction. While we do not claim that this study reflects the entire spectrum of factors that induce visual discomfort when viewing S3D (such as color mismatches and other rivalries), it certainly addresses important disparity-related aspects of discomfort. Beyond visual discomfort prediction, there are a large number of unsolved problems for evaluating or quantifying the quality of experience (QoE) of S3D content, such as immersion, presence, and sense of reality. Moreover, it is difficult to find studies directed towards understanding the

QoE of non-stationary S3D video content. Given the rapid growth in demand for S3D virtual reality, the aforementioned QoE problems need to be properly addressed towards reliably delivering high-quality visual content.

REFERENCES

[1] M. Emoto, T. Niida and F. Okano, "Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television," *Journal of Display Technology,* vol. 1, no. 2, pp. 328-340, December 2005.

[2] H. Sohn, Y. Jung, S. Lee and Y. Ro, "Predicting visual discomfort using object size and disparity information in stereoscopic images," *IEEE Trans. Broadcast.,* vol. 59, no. 1, pp. 28-37, March 2013.

[3] Y. Jung, H. Sohn, S. Lee and Y. Ro, "Visual comfort improvement in stereoscopic 3D display using perceptually plausible assessment metric of visual comfort," *IEEE Trans. Consumer Elec.,* vol. 60, no. 1, pp. 1-9, April 2014.

[4] M. Lambooij, W. Ijsselsteijn and I. Heynderickx, "Visual discomfort in stereoscopic displays: a review," *J. Imaging Sci. Technol.,* vol. 53, no. 3, pp. 030201.030201-14, May 2009.

[5] T. Bando, A. Iijima and S. Yano, "Visual fatigue caused by stereoscopoic images and the search for the requirement to prevent them: a review," *Displays,* vol. 33, no. 2, pp. 76-83, April 2012.

[6] D. M. Hoffman, A. R. Girshick, K. Akeley and M. S. Banks, "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *J. Vision,* vol. 8, no. 3, pp. 1-30, March 2008.

[7] H. Oh, S. Lee and A. C. Bovik, "Stereoscopic 3D visual discomfort prediction: A dynamic accommodation and vergence interaction model," *IEEE Trans. Image Process.,* vol. 25, no. 2, pp. 615-629, Feb. 2016.

[8] J. Park, H. Oh, S. Lee and A. C. Bovik, "3D visual discomfort predictor: Analysis of disparity and neural activity statistics," *IEEE Trans. Image Process.,* vol. 24, no. 3, pp. 1101-1114, March 2015.

[9] T. Kim, S. Lee and A. C. Bovik, "Transfer function model of physiological mechanisms underlying temporal visual discomfort experienced when viewing stereoscopic 3D images," *IEEE Trans. Image Process.,* vol. 24, no. 11, pp. 4335-4347, Nov. 2015.

[10] S. Yano, S. Ide, T. Mitsuhashi and H. Thwaites, "A study of visual fatigue and visual comfort for 3D HDTV/HDTV," *Displays,* vol. 23, no. 4, pp. 191-201, June 2002.

[11] S. P. Du, D. Masia, S. M. Hu and D. Gutierrez, "A metric of visual comfort for stereoscopic motion," *ACM Transactions on Graphics (TOG),* vol. 32, no. 6, pp. 222:1-222:9, 2013.

[12] S. Poulakos, R. Monroy, T. Aydin and O. Wang, "A computational model for perception of stereoscopic window violations," *IEEE International Workshop on Quality of Multimedia Experience (QoMEX),* pp. 1-6, 2015.

[13] Y. Nojiri, H. Yamanoue, A. Hanazato, and F. Okano, "Measurement of parallax distribution and its application to the analysis of visual comfort for stereoscopic HDTV," in *Proc. SPIE, Stereoscopic Displays Virtual Reality Syst. X,* vol. 5006, pp. 195-205, May 2003.

[14] J. Choi, D. Kim, S. Choi and K. Sohn, "Visual fatigue modeling and analysis for steroscopic video," *Opt. Eng.,* vol. 51, no. 1, pp. 017206.017206-11, Jan. 2010.

[15] D. Kim and K. Sohn, "Visual fatigue prediction for steroscopic image," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 21, no. 2, pp. 231-236, Feb. 2011.

[16] Y. Jung, H. Sohn, S. Lee, H. Park and Y. Ro, "Predicting visual discomfort of stereoscopic images using human attention model," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 23, no. 12, pp. 2077-2082, Dec. 2013.

[17] J. Park, S. Lee and A. C. Bovik, "3D visual discomfort prediction: vergence, foveation, and the physiological optics of accommodation," *IEEE J. Select. Topics Signal Process.,* vol. 8, no. 3, pp. 415-427, June 2014.

[18] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-444, May 2015.

[19] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Net,* vol. 61, pp. 85-117, Jan. 2015.

[20] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang and A. C. Bovik, "Deep convolutional neural models for picture quality prediction," *IEEE Signal Processing Magazine,* vol. 34, no. 6, pp. 103-141, Nov. 2017.

[21] IEEE-SA Stereo Image Database 2012 [online]. Available: http://grouper.ieee.org/groups/3dhf/

[22] Recommendation, ITU-R BT. 500-11, "Methodology for the subjective assessment of quality of television pictures," Standardization Sector of ITU, 2002.

[23] Y. Jung, H. Sohn, S. Lee, H. Park and Y. Ro, "Predicting visual discomfort of stereoscopic images using human attention model," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 23, no. 12, pp. 2077-2082, Dec. 2013.

[24] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2009, pp. 248-255.

[25] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Tech. Rep.,* 2009.

[26] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Select. Topics Signal Process.,* vol. 11, no. 1, pp. 206-220, Feb. 2017.

[27] H. Oh, J. Kim and S. Lee, "Blind deep S3D image quality evaluation via local to global feature aggregation," *IEEE Trans. Image Process.,* to appear..

[28] A. W. Roe, A. J. Parker, R. T. Born, and G. C. DeAngelis, "Disparity channels in early vision," *J. Neurosci.,* vol. 27, no. 44, pp. 11820-11831, Oct. 2007.

[29] G. C. DeAngelis, B. G. Cumming and W. T. Newsome, "Cortical area MT and the perception of stereoscopic depth," *Nature,* vol. 394, pp. 677-680, Aug. 1998.

[30] G. C. DeAngelis and T. Uka, "Coding of horizontal disparity and velocity by MT neurons in the alert macaque," *J Neurophysiol.,* vol. 89, pp. 1094-1111, 2003.

[31] M. Wildeboer, O. Stankiewicz and K. Wegner, "A semi-automatic multi-view depth estimation method," *Proceeding of SPIE Visual Communications and Image Processing,* vol. 7744, pp. 77442B-1, 2010.

[32] K. Umeda, S. Tanabe and I. Fujida, "Representation of stereoscopic depth based on relative disparity in macaque area V4," *J. Neurophysiol.,* Vol, 98, pp. 241-252, 2007.

[33] M. Clark and A. C. Bovik, "Experiments in segmenting texton patterns using localized spatial filters," *Pattern recognition,* Vol, 22, no. 6, pp. 707-717, 1989.

[34] E. Kandel, J. Schwartz and T. M. Jessel, *Principles of Neural Science,* Elsevier, New York, 1990.

[35] T. D. Sanger, "Neural population codes," *Curr. Opin. Neurobiol.,* vol. 13, pp. 238-249, 2003.

[36] R. S. Zemel, P. Dayan and A. Pouget, "Probabilistic interpretation of population codes," *Neural Comput.,* vol. 10, pp. 403-430, 1998.

[37] A. Maalouf and M. C. Larabi, "CYCLOP: a stereo color image quality assessment metric," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP),* May 2011, pp. 1161-1164.

[38] M. J. Chen, C. C. Su, D. L. Kwon, L. K. Cormack and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Commun.,* vol. 28, no. 9, pp. 1143-1155, Oct. 2013.

[39] D. Ruderman, "The statistics of natural images," *Network Computation in Neural Systems,* vol. 5, no. 4, pp. 517-548, 1994.

[40] K. Jarrett, K. Kavukcuoglu, M. Ranzato and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *IEEE Conf. Computer Vision (ICCV),* 2009, pp. 2146-2153.

[41] S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, "Data preprocessing for supervised learning," *Int. J. Computer Science,* vol. 1, no. 1, pp. 111-117, 2006.

[42] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Letter,* vol. 20, no. 3, pp. 209-212, March 2013.

[43] M. Saad, A. C. Bovik and C. Charrier, "Blind image quality assessment: a natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.,* vol. 21, no. 8, pp. 3339-3352, Aug. 2012.

[44] A. Mittal, A. Moorthy and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.,* vol. 21, no. 12, pp. 4695-4708, Dec. 2012.

[45] M. J. Chen, L. K. Cormack and A. C. Bovik, "No-reference quality assessment of natural stereopairs," *IEEE Trans. Image Process.,* vol. 22, no. 9, pp. 3379-3390, Sept. 2013.

[46] D. Kane, P. Guan and M. S. Banks, "The limits of human stereopsis in space and time," *J. Neuroci.,* vol. 34, no. 4, pp. 1397-1408, Jan. 2014.

[47] K. Lee, A. K. Moorthy, S. Lee and A. C. Bovik, "3D visual activity assessment based on natural scene statistics," *IEEE Trans. Image Process.,* vol. 23, no. 1, pp. 450-465, Jan. 2014.

[48] Y. Nojiri, H. Yamanoue, S. Ide, S. Yano and F. Okana, "Parallax distribution and visual discomfort on stereoscopic HDTV," in *Proc. IBC.,* pp. 373-380, Sept. 2006.

[49] H. Kim, S. Lee and A. C. Bovik, "Saliency prediction on stereoscopic videos," *IEEE Trans. Image Process.,* vol. 23, no. 4, pp. 1476-1490, Apr. 2014.

[50] A. L. Maas, A. Y. Hannun and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int'l Conf. Machine Learning (ICML),* 2013.
[51] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Machine Learning (ICML),* 2010, Haifa, Israel.
[52] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," *ArXiv150203167Cs,* 2015.
[53] I. Sutskever, J. Martens, G. Dahl and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. 30th Int. Conf. Machine Learning (ICML),* 2013, pp. 1139-1147.
[54] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. 32th Int. Conf. Machine Learning (ICML),* 2015.
[55] L. Kang, P. Ye, Y. Li and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2014, pp. 1733-1740.
[56] J. Park, K. Seshadrinathan, S. Lee and A. C. Bovik, "VQPooling: video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Process.,* vol. 22, no. 2, pp. 610-620, Feb. 2013.
[57] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Select. Topics Signal Process.,* vol. 3, no. 2, pp. 193-201, March 2009.
[58] C. W. Dunnett, "A multiple comparison procedure for comparing several treatments with a control," *J. American Statistical Associ.,* vol. 50, no. 272, pp. 1096-1121, 1955.
[59] H. Oh and S. Lee, "Visual presence: viewing geometry visual information of UHD S3D entertainment," *IEEE Trans. Image Process.,* vol. 25, no. 7, pp. 3358-3371, July 2016.
[60] T. Shibata, J. Kim, D. M. Hoffman and M. S. Banks "The zone of comfort: Predicting visual comfort with stereo displays," *J. Vision,* vol. 11, no. 8, pp. 1-29, July 2011.
[61] L. Goldmann, F. D. Simone and T. Ebrahimi, "Impact of acquisition distortions on the quality of stereoscopic images," in *Int'l Wkshp Video Process. Quality Metrics Consumer Elec. (VPQM),* 2010.
[62] D. Mishkin, N. Segievskiy and J. Matas, "Systematic evaluation of CNN advances on the ImageNet," *ArXiv1606.02228v2,* 2016.
[63] V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets and V. Lempitsky, "Speeding-up convolutional neural networks using fine-tuned CP-decomposition," in *Int'l Conf. Learning Representation (ICLR),* 2015.
[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int'l Conf. Learning Representation (ICLR),* 2015, pp. 1-14.
[65] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS),* 2012, pp. 1097-1105.
[66] H. Kim and S. Lee, "Transition of visual attention assessment in stereoscopic images with evaluation of subjective visual quality and discomfort," *IEEE Trans. Multimedia,* vol. 17, no. 12, pp. 2198-2209, Dec. 2015.
[67] Y. Liu, L. K. Cormack and A. C. Bovik, "Dichotomy between luminance and disparity features at binocular fixations," *J. Vision,* vol. 10, no. 12, pp. 1-17, Oct. 2010.
[68] L. V. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.,* vol. 9, no. 85, pp. 2579-2605, 2008.

Heeseok Oh received the B.S., M.S., and Ph.D. degrees in electrical and electronics engineering from Yonsei University, Seoul, South Korea, in 2010, 2012, and 2017, respectively, where he is currently with Samsung Electronics Company Ltd., Seoul, South Korea. His current research interests include 2D/3D image and video processing based on human visual system, machine learning, and computational vision.

Sewoong Ahn received the B.S. degree in fusion electronic engineering from Hanyang University, Seoul, South Korea, in 2015. He is currently pursuing the M.S. and Ph.D. degrees with the Multidimensional Insight Laboratory, Yonsei University. His research interests include 2-D/3-D image and video processing based on human visual system, quality assessment of 2-D/3-D image and video, 3-D virtual reality, and deep learning.

Sanghoon Lee (M'05–SM'12) received the B.S. in E.E. from Yonsei University in 1989 and the M.S. in E.E. from Korea Advanced Institute of Science and Technology (KAIST) in 1991. From 1991 to 1996, he worked for Korea Telecom. He received his Ph.D. in E.E. from the University of Texas at Austin in 2000. From 1999 to 2002, he worked for Lucent Technologies on 3G wireless and multimedia networks. In March 2003, he joined the faculty of the Department of Electrical and Electronics Engineering, Yonsei University, Seoul, Korea, where he is a Full Professor. He was an Associate Editor of the IEEE Trans. Image Processing (2010-2014). He has been an Associate Editor of IEEE Signal Processing Letters (2014-) and Journal of Electronic Imaging (2015-) and Chair of the IEEE P3333.1 Quality Assessment Working Group (2011-). He currently serves on the Technical Committee of the IEEE Multimedia Signal Processing (2016-) and the IEEE IVMSP Technical Committee (2014-), was Technical Program Co-chair of the International Conference on Information Networking (ICOIN) 2014, and of the Global 3D Forum 2012 and 2013, and was General Chair of the 2013 IEEE IVMSP Workshop. He also served as a special issue Guest Editor of IEEE Trans. Image Processing in 2013, and as Editor of the Journal of Communications and Networks (JCN) (2009-2015). He received a 2015 Yonsei Academic Award from Yonsei University, a 2012 Special Service Award from the IEEE Broadcast Technology Society and a 2013 Special Service Award from the IEEE Signal Processing Society. His research interests include image/video quality assessment, computer vision, graphics, cloud computing and multimedia communications and wireless networks.